THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 258, No. 6, Issue of March 25, pp. 3991-3995, 1983
Printed in U.S.A.

A Method for Mapping RNA Initiation, Termination, Splice, and Protein Binding Sites

RIBOSOME BINDING SITES ON β-GLOBIN MESSENGER RNA*

(Received for publication, September 27, 1982)

Jeffrey R. Patton and Chi-Bom Chae

From the Department of Biochemistry, University of North Carolina, Chapel Hill, North Carolina 27514

A new method for mapping RNA initiation, termination, and splice sites was developed. The method involves: 1) hybridization of RNA to end-labeled single-stranded DNA; 2) mild digestion of the hybrid with a single strand specific nuclease; 3) high resolution gel electrophoresis and autoradiography. The regions of the labeled probe which are resistant to nuclease digestion are mapped by measuring the distance from the labeled end. The sequence can be determined by running the end-labeled probe sequenced according to the protocol of Maxam and Gilbert (Maxam, A. M., and Gilbert, W. (1980) *Methods Enzymol.* 65, 499-560) at the same time. The feasibility of this method was tested with adult chicken β-globin mRNA, and we found that the transcription initiation, termination, and RNA splice sites can be determined to within a few bases of the known sites. Using this method we also found that ribosomes bind to β-globin mRNA from 22 bases upstream from the translation start codon (AUG) to 15 bases past the stop codon (UAA).

Several methods are currently available to map RNA initiation, termination, and splice sites. The transcription initiation site can be mapped by the S1 nuclease mapping method of Berk and Sharp (1, 2). In this method a DNA restriction fragment spans the reputed initiation site and has one 5′ end-labeled end in the first coding block. After hybridization with RNA and subsequent digestion with S1 nuclease the initiation site is determined from the length of protected DNA. A primer extension method can also be used to map the initiation site (3).

Uniformly labeled genomic DNA can be hybridized to mRNA and the hybrid digested with S1. From the size of the labeled DNA fragments generated, one can determine if the transcript from a gene inserted into an expression vector or a gene transferred into cells is the same as the naturally occurring RNA transcript (4). However, this method is not suitable for mapping transcription initiation, termination, and splice sites of unknown RNA.

Protein binding sites on DNA are usually mapped by the DNase I footprinting method of Galas and Schmitz (5). In this method, purified proteins are bound to an end-labeled restriction fragment of DNA and mildly digested with DNase I. The region protected from DNase I digestion appears as a blank area on the autoradiograph of a denaturing gel. A similar approach can be used for mapping protein binding sites on RNA, but unless the RNA being studied is small, it is difficult to generate unique lengths of end-labeled RNA which can be resolved by gel electrophoresis. Sequencing the RNA fragments protected from nuclease digestion by bound proteins is another approach to mapping protein binding sites on RNA (6). In this case one has to purify a homogeneous piece of RNA in order to sequence the fragment.

In this report we describe a method which overcomes these difficulties in mapping RNA initiation sites, termination sites, and splice sites. In addition, the method can also be applied to mapping the nuclease-resistant RNA fragments generated from RNA-protein complexes. To test the feasibility of this method we mapped the transcription initiation and termination sites and splice sites of chicken β-globin RNA. Also the ribosome binding sites on the β-globin mRNA in chicken reticulocyte polyribosomes were determined.

MATERIALS AND METHODS

Isolation of Polysomes—Erythroid cells were obtained from white leghorn chickens after phenylhydrazine treatment and polysomes were isolated as described (7).

Preparation of Polysomal and Protected Polysomal RNA—To prepare undegraded total polysomal RNA the polysomal pellet was suspended in 1% SDS,[1] 10 mM Tris, pH 7.5, 1 mM EDTA, 0.15 M NaCl. Vadanyl-adenosine complex (VSA) was added to the solution as a ribonuclease inhibitor (8). The mixture was treated with proteinase K (50 μg/ml) for 2 h at 37 °C and extracted 3 times with phenol (saturated with 10 mM Tris, pH 7.5, 1 mM EDTA) and once with $CHCl_3$. The RNA was precipitated with 2.5 volumes of ethanol after making the solution 0.3 M sodium acetate.

To prepare micrococcal nuclease-protected polysomal RNA the polysomal pellet was taken up in 5% glycerol, 10 mM Tris, pH 7.5, 0.2 mM $MgCl_2$, 1 mM $CaCl_2$, and insoluble material was removed by a short centrifugation (15 min, 12,000 × *g*, 4 °C). The supernatant was adjusted to a nucleic acid concentration of 1 mg/ml and digested with micrococcal nuclease (1.5 μg/ml) for 2 h at 22 °C. After digestion the solution was made 1 mM EDTA, 1% SDS and treated with proteinase K (50 μg/ml), in the presence of vadanyl-adenosine complex for 2 h at 37 °C. The solution was extracted with phenol and the RNA precipitated with ethanol.

Cloning—The chicken β-globin gene fragments that will be described under "Results" were inserted into M13mp7RF by established methods and with observation of current NIH guidelines for recombinant DNA research. The sources of the β-globin gene fragments were a λ phage containing adult β-globin genomic DNA (λCβG1) supplied by Dodgson *et al.* (9) and an adult β-globin cDNA plasmid (pHB1001) supplied by Salser *et al.* (10).

Preparation of Single-stranded End-labeled DNA Fragments—The various single-stranded globin gene fragments used in this investigation were isolated from the single-stranded recombinant phage DNA of M13mp7 by digestion with *Bam*H1 as described previously

* This work was supported by National Institutes of Health Grant GM 27839. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] The abbreviations used are: SDS, sodium dodecyl sulfate; P-RNA, nuclease-protected polysomal RNA; M13mp7RF, replicative form of M13mp7 DNA.

(12). The isolated fragments were treated with bacterial alkaline phosphatase (Worthington) and labeled with ^{32}P at the 5′ end with T4 polynucleotide kinase (Bethesda Research Laboratories) according to the method of Maxam and Gilbert (11).

Hybridization—The hybridization of RNA to the end-labeled probes was carried out in 0.4 M NaCl, 10 mM Tris, pH 7.5, 1 mM EDTA at 68 °C for >16 h. Usually 50,000 dpm of end-labeled probe (5–20 ng of β-globin gene) and 50 μg of RNA were mixed in a final volume of 25 μl.

Digestion of Hybrids with Mung Bean Nuclease—The hybridization reactions were taken up in 1 ml of 30 mM sodium acetate, pH 4.6, 50 mM NaCl, 1 mM $ZnCl_2$, 5% glycerol and equilibrated at 37 °C. Mung bean nuclease (6.5 units, P-L Biochemicals) was added, and the solution was incubated for 30 s. The reaction was stopped with EDTA (final concentration 16 mM), and the solutions were made 0.3 M sodium acetate and precipitated with 2.5 volumes of ethanol at −20 °C for >16 h. After centrifugation (12,000 × *g*, 20 min, 4 °C) the pellets were treated with 0.3 N NaOH at 68 °C for 1 h. The reaction was neutralized with HCl and Tris, pH 7.5, yeast tRNA was added as carrier (3 μg), and the solution was precipitated with 2.5 volumes of ethanol at −20 °C for >16 h. After centrifugation (12,000 × *g*, 15 min, 4 °C) the pellet was dried under vacuum and taken up in 5 μl of tracking dye (80% formamide, 1 mM EDTA, 10 mM NaOH, 0.1% xylene cyanol, 0.1% bromphenol blue). Equal amounts of cold acid-precipitable radioactivity were loaded on the gel.

Sequencing—The procedure of Maxam and Gilbert (11) was used to sequence the 5′ end-labeled fragments.

Acrylamide Gel Electrophoresis—Sequencing gels (0.3, 330, and 400 mm) were prepared as described by Maxam and Gilbert (11). The gel was exposed to Kodak XAR-5 x-ray film with a DuPont Cronex intensifying screen at −70 °C.

RESULTS AND DISCUSSION

The basic strategy used in mapping RNA is similar to that of the DNase I footprinting method of Galas and Schmitz (5). The RNA to be mapped is hybridized to an end-labeled single-stranded DNA fragment and the hybrid is mildly digested with a single strand specific nuclease, such as S1 or mung bean nuclease, in such a way that the enzyme cuts the end-labeled DNA approximately once per molecule. High resolution gel electrophoresis and autoradiography will show the DNA bands containing the labeled end. The regions of the probe that form a RNA-DNA hybrid will not be cut by the enzyme and should appear as blanks, and the regions that do not form a hybrid will show ladder patterns. The region protected by RNA can be mapped by measuring the distance from the labeled end. Also by running the end-labeled DNA cleaved by the base-specific reagents of Maxam and Gilbert (11) concomitantly, one can deduce the sequence of the region protected by RNA. In this report we mapped the transcription initiation and termination sites and the RNA splice sites in chicken β-globin RNA. Chicken reticulocyte polysomal RNA was used as the source of β-globin mRNA. Reticulocyte polysomes were digested with micrococcal nuclease and the nuclease-resistant RNA was isolated for mapping ribosome binding sites on β-globin mRNA.

Single-stranded probes covering the β-globin gene regions were isolated from the recombinant phage DNA of M13mp7. The areas cloned into M13mp7 are outlined in Fig. 1. The 406-base pair *Alu*I-*Pvu*II genomic fragment containing the first coding block was inserted into the *Sal*I site of M13mp7RF DNA after adding *Sal*I linkers (mβG6). The 600-base pair *Sst*I genomic fragment containing the third coding block, was also inserted into the *Sal*I site by blunt-end ligation after trimming the ends of the DNAs (mβG7). mβG2 is a recombinant containing the 466-base pair *Hpa*II fragment of adult β-globin cDNA plasmid pHB1001 (10) inserted into the *Acc*I sites of M13mp7RF. The DNA inserts are in an orientation such that the phage DNA contains the strand complementary to mRNA. The single-stranded globin inserts were isolated from the single-stranded phage DNA by digestion with

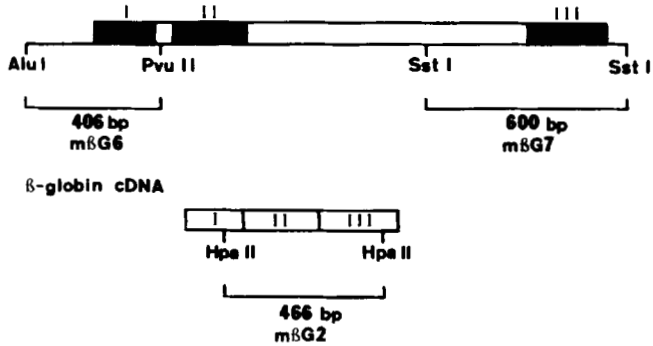


FIG. 1. **The chicken β-globin gene fragments used as probes.** bp, base pairs.

*Bam*H1 restriction endonuclease as described before (12). The fragments were labeled at the 5′ end with T4 polynucleotide kinase and [γ-^{32}P]ATP (11). The labeled probe and RNA were hybridized as described under "Materials and Methods," with the concentrations of globin RNA in at least 100-fold excess over the probe. The substrate concentration (RNA and probe) for mung bean nuclease was the same in all the samples to obtain comparable digestion of the probe in all the samples. The control sample contained an equal amount of tRNA. We found that mung bean nuclease produces more bands than S1 nuclease with less double-stranded endonuclease activity. Mung bean nuclease is very susceptible to SDS, however, and SDS was not included in the hybridization mixture. The concentration of mung bean nuclease and time of digestion were adjusted in such a way that a significant amount of uncut probe remains after digestion.

The RNA initiation site and the first splice donor site were mapped with the 430-base fragment (mβG6). This DNA contains the 406-base region which includes 207 bases of the 5′ flanking region, the first coding block (173 bases), and 26 bases of the first intron. The extra DNA (24 bases) was derived from the linkers used and M13mp7. In Fig. 2 the first splice junction was determined by sequencing the fragment. Having just the cytosine and thymine sequence ladder was found to be sufficient to determine the sequence of the splice donor since the sequence of β-globin mRNA is known (13). The DNA sequences of the chicken genomic β-globin gene and the RNA splice site are not known. By comparing the mRNA sequence (13) and the sequence of the 430-base probe (figures not shown here) we found that the RNA splice site is at the same position as the site seen with mammalian β-globin mRNA (9, 14). We determined that the first splice is 173 bases from the CAP site. The figure shows that polysomal RNA (*lane 1*) protects a stretch of 173 bases of DNA, and this region is bounded by ladder patterns. The junctions at the protected and unprotected regions correspond to the RNA initiation site (*CAP*) and the first splice donor site (*IVS*). The splice donor site can be mapped to within 3 bases of the actual site. However, the last base of the first exon and two bases in the intron, as determined by sequencing, are protected from nuclease attack by polysomal RNA (*lane 1*). To improve resolution in the region containing the initiation site the gel was run for an extended period, along with the probe treated with the base-specific reagents (Fig. 3). The position of the initiation site was found to be within 2 bases of the known initiation site (14). The only ambiguity in determining the boundary between the protected and unprotected regions in mapping the initiation and the splice sites is the uneven digestion of the end-labeled probe by the single strand specific nuclease. As can be seen in the control lane (see *lane 2* in Fig. 2) mung bean nuclease does not necessarily cut at every base,

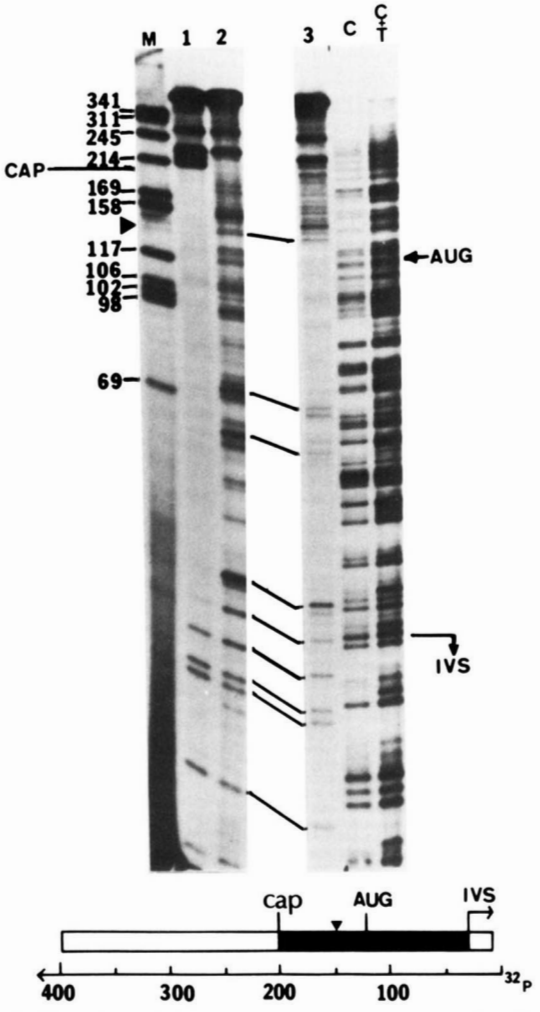


FIG. 2. **Mapping of the transcription initiation site, the first splice donor site, and the start of ribosome binding site of chicken β-globin RNA.** The 430-base probe from mβG6 containing the first coding region of the chicken β-globin gene (see *scale map* at *bottom*) was hybridized with total polysomal RNA (*lane 1*), tRNA (*lane 2*, control), and P-RNA (*lane 3*, polysomes treated with micrococcal nuclease) and digested with mung bean nuclease. *M*, molecular weight markers, the size is shown on the *left* in bases; *C*, the probe cleaved at cytosine residues; *C+T*, the probe cleaved at cytosine and thymine; *CAP*, transcription initiation site; *AUG*, translation start codon; *IVS*, the first splice donor site; ◀, the start of ribosome binding site. A scale map of the probe and the position of the label are shown at the *bottom*.

and there are regions where the enzyme cuts at a much slower rate. Therefore, depending on the sequence of the probe there will be a certain amount of error in determining the initiation and splice sites. However, the maximum error does not exceed 2–3 bases in the case of β-globin RNA. The S1 mapping method of Berk and Sharp (1) also has a certain amount of error due to the tendency of S1 to digest the end of duplex molecules.[2]

The ribosome binding sites in the first coding region were also mapped with P-RNA. As seen in Figs. 2 and 3, P-RNA protects the probe from 22 bases upstream (*arrowhead*) from the translation start codon, AUG. From this point until the start of the intervening sequence the lane does not show a clear blank as does the lane with undegraded polysomal RNA. Instead there are bands that also appear in the control lane but are diminished in intensity. One explanation for this observation is that polysomes contain a random population of ribosomes which have progressed to different parts of the mRNA during translation. Thus, P-RNA may represent a complex mixture of globin mRNA fragments which are derived from the entire region of the globin RNA between the start and end of the ribosome binding sites. P-RNA should produce the same DNA ladder pattern as the control but reduced in intensity, compared to the control, between the start and end of the ribosome binding sites. P-RNA protects the probe 4 bases short of the splice donor site, for some unknown reason.

Fig. 4 shows the result obtained with the 490-base probe containing the 466-base globin cDNA fragment (mβG2). As expected polysomal RNA shows a clear blank except at the extreme ends which contain M13mp7 sequences. When P-RNA was included in the hybridization, the probe was protected until 15 bases past the translation stop codon, UAA (*arrowheads*). This region does not show a clear blank like the lane with polysomal RNA. In the second coding region and near the 5′ end of the third coding region there are sites which are not protected by ribosomes (*asterisks*). These sites were mapped, and they occur at approximately 255, 185, 160, 145, and 107 bases from the labeled 5′ end. These could be in regions where ribosomes traverse faster than other regions. The region in the probe indicated by a *bracket* is relatively resistant to mung bean nuclease. This region spans almost 50

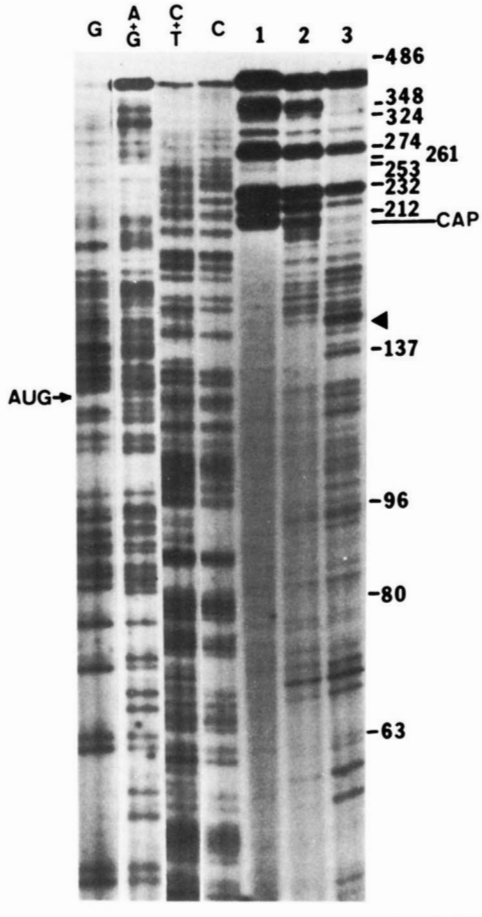


FIG. 3. **Detailed mapping of the transcription initiation site and the start of ribosome binding site.** The same probe and RNAs were used as in Fig. 2. The samples were electrophoresed longer to get better sequence information in the region of the transcription initiation site. *Lane 1*, polysomal RNA; *lane 2*, P-RNA; *lane 3*, tRNA control. *G*, *A+G*, *C+T*, and *C* are base specific sequencing reactions. *CAP*, *AUG*, and ◀ indicate the transcription initiation site, the translation start codon, and the start site of ribosome binding, respectively.

[2] P. Sassone-Corsi and P. Chambon, unpublished results.

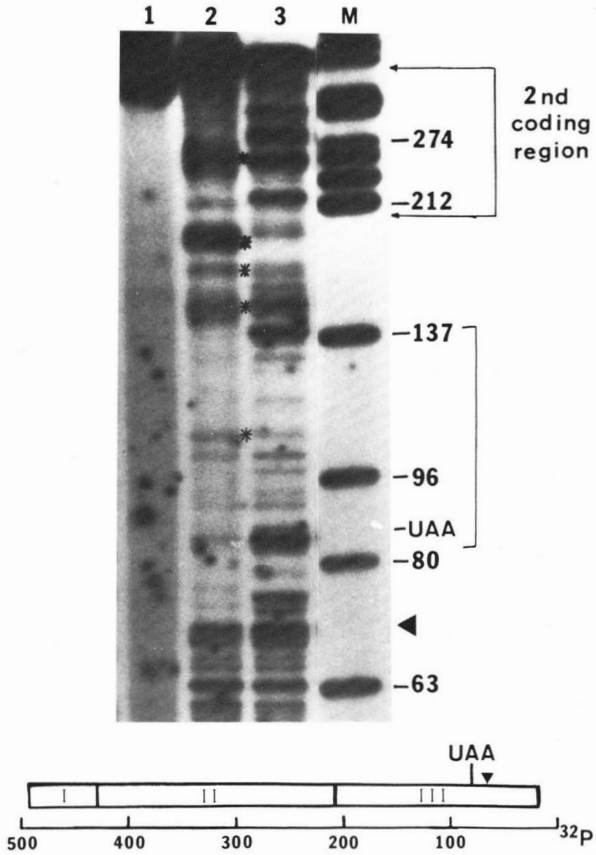


FIG. 4. **Mapping with the cDNA probe.** The 490-base cDNA probe (see *scale map* at *bottom*) was hybridized with polysomal RNA (*lane 1*), P-RNA (*lane 2*), and tRNA (*lane 3*, control). *M*, molecular weight markers with the sizes shown on the *right*. Only the relevant area of the gel is shown. The sequence reactions were run at the same time but not shown here. The *bracket* indicates the area in the probe which is relatively resistant to mung bean nuclease. The end of ribosome binding is denoted by ◀; *, regions not protected by ribosomes against micrococcal nuclease digestion.

bases and most likely forms a folded structure. The same region is shown with another probe (see Fig. 5).

The transcription termination site and the last splice acceptor site were mapped with the fragment containing the third coding block. The 600-base fragment (mβG6) contains approximately 300 bases of intervening sequences, the third coding sequence, and 58 bases of the 3′ untranscribed region. The results obtained with this fragment are shown in Fig. 5. Polysomal RNA (*lane 1*) shows a region which is a clear blank and this corresponds to the length of the third coding region. The boundary between the digested and undigested regions in *lane 1* corresponds to the splice acceptor site (*IVS*) and the poly(A) addition site (*POLY A*). The splice acceptor site was found to be within 2 bases of the putative splice site (at 299 instead of 297 bases). The splice acceptor site was determined from the sequence of mRNA (13) and that of the 600-base probe. Again we find that the second splice site occurs at the same position as the site in mammalian β-globin mRNA (9, 14). The poly(A) addition site agrees exactly with the published site for β-globin RNA (13).

P-RNA (*lane 2*) shows an area of reduced digestion by mung bean nuclease and the boundaries between the reduced and unreduced areas are the splice junction (*IVS*) and the junction indicated by an *arrowhead*. This site is again 15 bases past the translation stop codon, UAA, and is in good agreement with the result obtained with the cDNA probe (see Fig. 4). The bands indicated by *asterisks* are the same sites seen in Fig. 4 which are not well protected by ribosomes against micrococcal nuclease digestion. The region indicated by a *bracket* is the region in the single-stranded probe which is relatively resistant to cleavage by mung bean nuclease. Overexposure of the same region in the cDNA probe is shown in Fig. 4, *lane 3*. There are other regions in the probe which are relatively resistant to mung bean nuclease presumably due to formation of a folded structure.

As demonstrated in this report the method used can map transcription initiation, termination, and RNA splice sites within a few bases of the reported sites. Also the method can be used to determine the protein binding sites on a known RNA (abundant or less abundant) as long as appropriate gene fragments are available. The same method may also be used for mapping the secondary structure of a known RNA, using S1 resistant RNA fragments. The main advantages of this

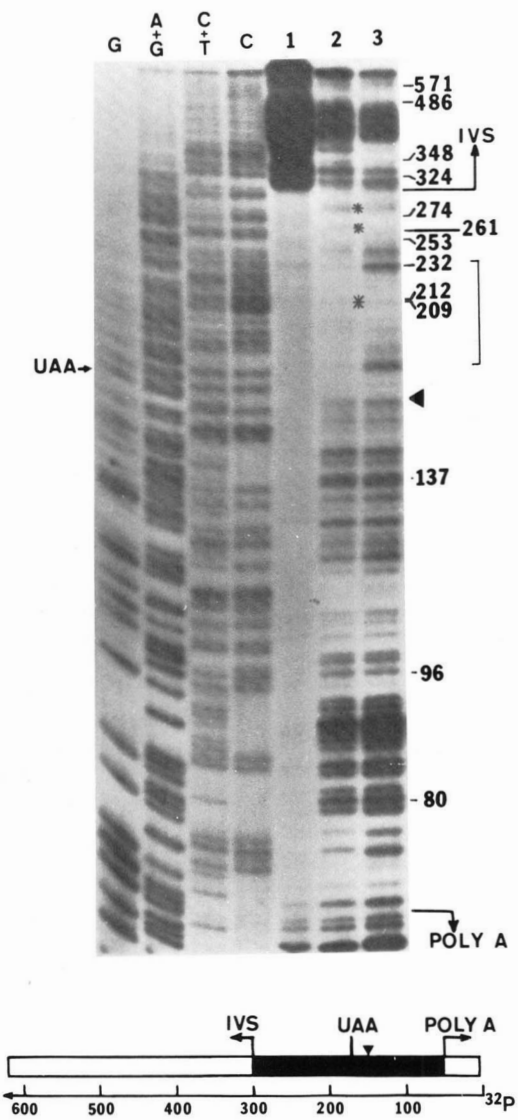
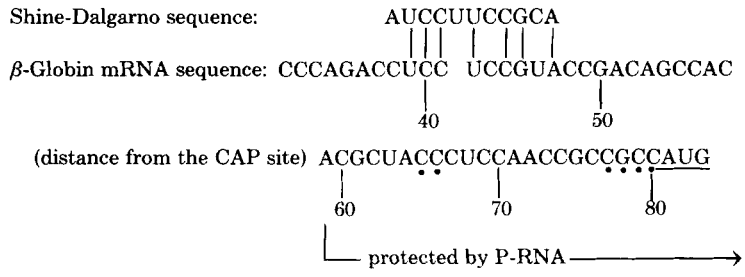


FIG. 5. **Mapping of the transcription termination site, the last splice acceptor site, and the end of ribosome binding in chicken β-globin RNA.** The 612-base probe (mβG7) containing the third coding region (see *scale map* at *bottom*) was hybridized with polysomal RNA (*lane 1*), P-RNA (*lane 2*), and tRNA (*lane 3*). *G*, *A+G*, *C+T*, and *C* are base specific sequencing reactions. The position of size markers are shown on the *right*. *UAA*, translation stop codon; *IVS*, the splice acceptor site, ◀, end of ribosome binding; *POLY A*, the poly(A) addition site. The scale map of the probe and the position of label are shown at the bottom.



SCHEME 1

method over the S1 mapping method of Berk and Sharp is that the position of the labeled end does not have to be within the RNA region. Also both sides of a splice site and the termination site can be determined easily. The major drawbacks of the method described here are: 1) that hybridizable RNA should be in excess over the probe, 2) that single strand specific nucleases do not produce bands at every base, and 3) that there are regions in the single-stranded probe that are relatively resistant to cleavage by mung bean nuclease, probably due to secondary structure. Therefore, it is important to include a control sample which contains no RNA that hybridizes to the probe to ensure that the boundary between nuclease-sensitive and nonsensitive regions does not fall into a nuclease-resistant region of the probe. However, the experiments with chicken β-globin mRNA show that the boundaries can be mapped within a few bases from the known initiation, termination, and RNA splice sites. The hybridization of RNA to the probe may disrupt the secondary structure of the single-stranded DNA and the nuclease may cut the end of duplex regions. If this is true there will be no difficulty in finding the boundary with reasonable accuracy.

The results using RNA derived from the polysomes digested with micrococcal nuclease show the ribosomes bind to the chicken β-globin mRNA starting at 22 bases upstream from the translation start codon, AUG. Also there is no ribosomal binding after 15 bases past the stop codon, UAA. The start site we determined is 11 bases away from the sequence which shows strong homology with the putative ribosome binding site, the Shine-Dalgarno sequence (15) (see Scheme 1). Day *et al.* (14) proposed another possible ribosome binding site closer to the start codon (indicated by dots) which shows weaker sequence homology with the putative ribosome binding site, the 3′ end of 18 S ribosomal RNA. Our result suggests that the site closer to the start codon may in fact be the ribosome binding site (Scheme 1). It has been shown that ribosomes protect about 35 bases of mRNA from nucleases (16, 17). The protection of 15 bases past the stop codon may be due to physical hindrance to nuclease attack by the ribosome bound to the UAA codon.

Acknowledgment—We thank Dr. David A. Ross for discussion of the mapping strategy, Dr. Clyde A. Hutchison, III, for the use of computer programs and advice on sequencing, Dr. Jerry Dodgson for the recombinant phage λCβG1, and Dr. Winston Salser for the cDNA plasmid pHB1001.

REFERENCES

1. Berk, A. J., and Sharp, P. A. (1977) *Cell* **12,** 721–732
2. Weaver, R. F., and Weissman, C. (1979) *Nucleic Acids Res.* **7,** 1175–1193
3. Baralle, F. E. (1977) *Nature* **267,** 279–281
4. Hamer, D. H., Kaehler, M., and Leder, P. (1980) *Cell* **21,** 697–708
5. Galas, D. J., and Schmitz, A. (1978) *Nucleic Acids Res.* **5,** 3157–3170
6. Kozak, M. (1977) *Nature* **269,** 390–394
7. Gadski, R. A., and Chae, C.-B. (1978) *Biochemistry* **17,** 869–874
8. Berger, S. L., and Birkenmeier, C. S. (1979) *Biochemistry* **18,** 5143–5149
9. Dodgson, J. B., Strommer, J., and Engel, J. D. (1979) *Cell* **17,** 879–887
10. Salser, W. A., Cummings, I., Liu, A., Strommer, J., Padayatty, J., and Clark, P. (1979) in *Cellular and Molecular Regulation of Hemoglobin Switching* (Stamatoyannopoulos, G., and Nienhuis, A., eds) pp. 621–643, Grune and Stratton, New York
11. Maxam, A. M., and Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560
12. Patton, J. R., and Chae, C.-B. (1982) *Anal. Biochem.* **126,** 231–234
13. Richards, R. I., Shine, J., Ullrich, A., Wells, J. R. E., and Goodman, H. M. (1979) *Nucleic Acids Res.* **7,** 1137–1146
14. Day, L. E., Hirst, A. J., Lai, E. C., Mace, M., and Woo, S. L. C. (1981) *Biochemistry* **20,** 2091–2098
15. Shine, J., and Dalgarno, L. (1974) *Proc. Natl. Acad. Sci. U. S. A.* **71,** 1342–1346
16. Kozak, M. (1981) in *Curr. Top. Microbiol. Immunol.* **93,** 81–123
17. Kozak, M., and Shatkin, A. J. (1978) *J. Biol. Chem.* **253,** 6568–6577